Claims

What is claimed is:

1. A method of determining the language of a textual passage, the method comprising the steps of:

(a) parsing said textual passage into a plurality of n-grams;

(b) comparing each of said n-grams with a plurality of databases, wherein each of said databases comprises a list of n-grams associated with a specific language;

(c) determining an initial weight for each of said n-grams, per language, by calculating the frequency with which each of said n-grams appears in each of said databases and dividing said frequency by the total number of n-grams in said respective database;

(d) determining the number of said databases within which each of said n-grams appear;

(e) altering said initial weight for each of said n-grams by multiplying said initial weight with the inverse of said number of databases within which each of said n-grams appear;

(f) producing the weight of each language over the text passage by calculating, per language, the sum over each n-gram in the text passage of the products of the number of times that that n-gram appears in the text passage and the language-

specific altered weight calculated in step (e) for that n-gram;

(g) sorting the list of per language passage weights from step (f) in decreasing order, returning the most likely language for the text passage as the first element (highest weight) in the list.

2. The method of claim 1 wherein the step of determining an initial weight for each of said n-grams, per language, comprises the steps of calculating the frequency with which each of said n-grams appears in each of said databases and dividing said frequency by the total number of n-grams in said respective database.

3. The method of claim 1 wherein said n-grams have a size selected from the group consisting of bi-grams, tri-grams, and quad-grams.

4. The method of claim 1 wherein said n-grams are anchored n-grams.

5. The method of claim 1 wherein said n-grams are replacement-type n-grams.

6. The method of claim 1 wherein said n-grams are any combination of n-grams, including anchored n-grams

and/or replacement-type n-grams, and/or n-grams of different lengths.

7. The method of claim 1 wherein said textual passage comprises 20 or more words.

8. The method of claim 1 wherein said textual passage comprises 40 or more words.

9. method of determining the language of a textual passage, the method comprising the steps of:

(a) filtering a plurality of short words from a textual passage;

(b) comparing each of said short words against a plurality of databases, wherein each of said databases comprises a list of short words associated with a different language;

(c) determining an initial weight for each of said short words, per language, by calculating the frequency with which each of said short words appears in each of said databases and dividing said frequency by the total number of short words in said respective database;

(d) determining the number of said databases within which each of said short words appear;

(e) altering said initial weight for each of said short words by multiplying said initial weight

with the inverse of said number of databases within which each of said short words appear;

producing the weight of each language over the text passage by calculating, per language, the sum over each short word in the text passage of the products of the number of times that that short word appears in the text passage and the language-specific altered weight calculated in step (e) for that short word;

(g) sorting the list of per language passage weights from step (f) in decreasing order, returning the most likely language for the text passage as the first element (highest weight) in the list.

10. A method of determining the language of a textual passage, the method comprising the steps of:

(a) filtering a plurality of short words from a textual passage and parsing said textual passage into a plurality of n-grams;

(b) comparing each of said n-grams and said short words against a plurality of databases, wherein each of said databases comprises a list of n-grams and short words associated with a different language;

(c) determining an initial weight for each of said n-grams and said short words, per language;

(d) determining the number of said databases within which each of said n-grams and said short words appear;

(e) altering said initial weight for each of said n-grams and said short words by multiplying said initial weight with the inverse of said number of databases within which each of said n-grams and said short words appear;

producing the weight of each language over the text passage by calculating, per language, the sum over each short word and each n-gram in the text passage of the products of the number of times that that short word or n-gram appears in the text passage and the language-specific altered weight calculated in step (e) for that short word or n-gram;

(g) sorting the list of per language passage weights from step (f) in decreasing order, returning the most likely language for the text passage as the first element (highest weight) in the list.

11. A system for determining the language of a textual passage, comprising:

a central processing unit coupled to a memory system and a display , wherein said central processing unit operates according to a program retrieved from said memory system, wherein said program includes the steps of;

(a) receiving a textual passage;

(b) parsing said textual passage into a plurality of n-grams;

(c) comparing each of said n-grams against a plurality of databases, wherein each of said databases comprises a list of n-grams associated with a different language;

(d) assigning an initial weight to each of said n-grams, per language, by calculating the frequency with which each of said n-grams appears in each of said databases and dividing said frequency by the total number of n-grams in said respective database;

(e) calculating the number of said databases within which each of said n-grams appear;

(f) altering said initial weight assigned to each of said n-grams by multiplying said initial weight with the inverse of said number of databases within which each of said n-grams appear;

(g) producing the weight of each language over the text passage by calculating, per language, the sum over each n-gram in the text passage of the products of the number of times that that n-gram appears in the text passage and the language-specific altered weight calculated in step (f) for that n-gram;

(h) sorting the list of per language passage weights from step (g) in decreasing order, returning the

most likely language for the text passage as the first element (highest weight) in the list.

12. The system of claim 11 further comprising a scanner and an optical character recognition device, wherein said scanner and said optical character recognition device are connected to said central processing unit, wherein said program receives a textual passage from a document scanned by said scanner.

13. The system of claim 11 wherein said program comprises a user interface that allows a user to enter said textual passage.

14. The system of claim 13 wherein said user interface is a graphical user interface.

15. The system of claim 13 wherein said user interface displays the identified language.

16. The system of claim 11 wherein said program comprises a user interface that allows a user to enter a Uniform Resource Locator that contains said textual passage.